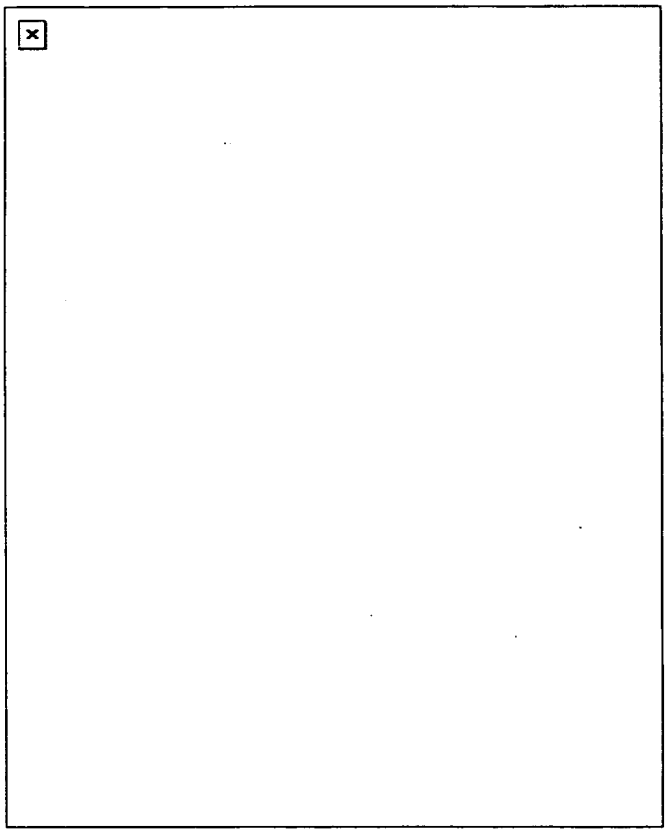
A7

DECISION TREE LEARNING AND GENERATING DEVICE

Patent number: JP7093158
Publication date: 1995-04-07
Inventor: ORIHARA RYOHEI
Applicant: TOSHIBA CORP
Classification:
- international: G06F9/44; G06F15/18
- european:
Application number: JP19930238150 19930924
Priority number(s):

Abstract of **JP7093158**

PURPOSE:To automatically generate a decision tree for feature amount classification to take continuous values.
CONSTITUTION:As the device to obtain rules for classifying a case data (DT) set into correspondent classification classes (CL) among previously decided CL corresponding to the plural kinds of feature amounts concerning the set of plural DT provided with the two kinds of feature amounts of a first feature amount (D1) expressing the feature of a case and a second feature amount (D2) expressing an attribute for classification at least, this device is constituted by providing a means to successively fetch the plural applied DT, to calculate difference (df) between the maximum value and minimum value of D1 by attributes, to divide it into a certain number of calculate decided by the df and relevant grading until the value of this df after division gets less than the value of grading of classification decided for classifying D1, to define the result as CL and to generate the decision tree by deciding the critical value range of the feature values of respective adjacent CL with the intermediate value of the closest value among the feature values of respective CL in the case of division, and a means to finish the generation of the decision tree when the class division error rate of respective CL reaches a previously decided set value.



(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平7－93158

(43)公開日　平成７年(1995)４月７日

(51)Int.Cl.[6]		識別記号	庁内整理番号	ＦＩ	技術表示箇所
Ｇ０６Ｆ	9/44	５５０　Ｍ	9193－5B		
	15/18	５６０　Ａ	9365－5L		

審査請求　未請求　請求項の数１　ＯＬ　（全 16 頁）

(21)出願番号　特願平5－238150

(22)出願日　平成５年(1993)９月24日

(71)出願人　000003078
株式会社東芝
神奈川県川崎市幸区堀川町72番地

(72)発明者　折原　良平
神奈川県川崎市幸区柳町70番地　株式会社東芝柳町工場内

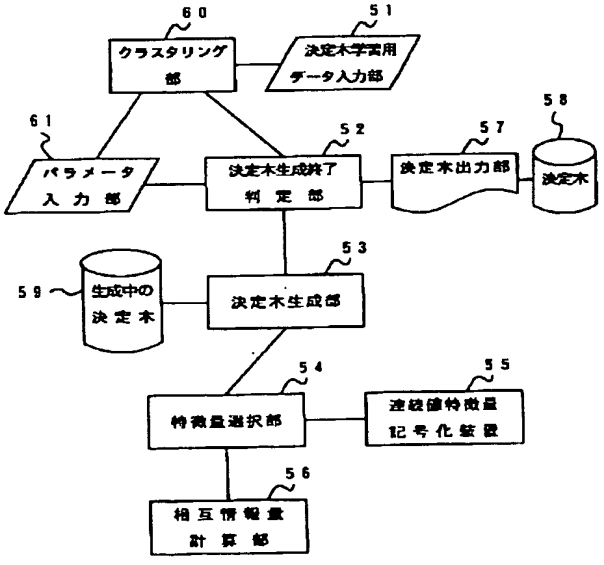
(74)代理人　弁理士　鈴江　武彦

(54)【発明の名称】　決定木学習生成装置

(57)【要約】

【目的】連続値をとる特徴量分類用の決定木を自動生成できるようにする。

【構成】事例の特徴を表す第１特徴量(D1)と、分類用の属性を示す第２特徴量(D2)との少なくとも２種の特徴量を有する事例データ(DT)を対象とし、複数のDTの集合に対し、前記複数種の特徴量に応じてそのDT集合を、予め定めた分類クラス(CL)のうちの対応するCLに分類するための規則を得る装置として、与えられた複数のDTを順次取り込み、そのD1の最大値と最小値との差(df)を属性別に求めこのdfの分割後の値がD1を分類するために定めた分類の粒度の値以下となるまでdfと当該粒度の値とにより定まるクラスタ数分に分割してCLとすると共に、各隣接するCLの特徴値の境界値範囲を前記分割の際の各CLの特徴値のうちの最も近い値の中間値を以て定めることで決定木を生成する手段、CL夫々のCLのクラス分け誤り率が予め定めた設定値に達すると決定木生成を終了する手段とを設けて構成した。



1

【特許請求の範囲】

【請求項１】　事例の特徴を表す第１の特徴量と、分類用の属性を示す第２の特徴量との少なくとも２種の特徴量を有する事例データを対象とし、複数の事例データの集合に対し、前記複数種の特徴量に応じてその事例データ集合を、予め定めた分類クラスのうちの対応する分類クラスに分類するための規則を得る決定木学習生成装置として、

与えられた複数の事例データを順次取り込み、その第１の特徴量の最大値と最小値との差を属性別に求め、この差の分割後の値が前記第１の特徴量を分類するために予め定めた分類の粒度の値以下となるまで、前記差と当該粒度の値とにより定まるクラスタ数分に分割して分類クラスとすると共に、各隣接する分類クラスの特徴値の境界値範囲を前記分割の際の各分類クラスの特徴値のうちの最も近い値の中間値を以て定めることにより決定木を生成する手段、分類クラスそれぞれの分割クラスのクラス分け誤り率が予め定めた設定値に達すると決定木生成を終了する決定木生成終了制御手段とを設けて構成したことを特徴とする決定木学習生成装置。

【発明の詳細な説明】

【０００１】

【産業上の利用分野】本発明は、与えられたいくつかの事例データを、適当な部分クラスに分類する分類規則を、事例データから帰納的に学習する決定木学習装置に関する。

【０００２】

【従来の技術】近年、人口知能の研究において、例題を与えることによって、対象とする概念の一般知識を発見する機械学習の技術が広く研究されつつある。その一つの方向として、数値または記号で表現された幾つかの特徴量と、それに対してオペレータ等が事前にデータを吟味して与える分類結果を事例とし、多数の事例を分類装置の例題として与えることによって、特徴量と分類結果の間の一般法則を見付け出し、新たな事例が与えられた時に、その事例の持つ特徴量を評価して分類結果を出力すると云った処理をする学習装置が開発されている。

【０００３】ここで、事例が分類されるべき部分クラスの分類結果を分類クラスと呼ぶ。例えば、図７に示す事例データは、２つの特徴量N_1（＝604 ～ 695）およびN_2（＝Ａ，Ｂ）と、分類クラス（Ｃ1 ～Ｃ5 ）とを持つ事例データである。

【０００４】そして、事例データの持つ特徴量を評価してその事例データの分類結果を得ると云ったような分類規則を導く方法の代表例として、特徴量を分類属性とし、分類結果クラスを得る決定木（「特徴量Ａの値がａであり、特徴量Ｂの値がｂであれば分類結果はＣである」と云った規則を、特徴量を“節”とし、分類結果を“葉”とする木（ツリー）の形にまとめたもの）を導くようにする手法がある。

2

【０００５】図８は、上述した図７の事例に対して、以下で説明する公知の決定木学習アルゴリズムが導いた決定木の例を示した図である。この決定木は、「特徴量N_2の値が“Ａ”であって、特徴量N_1の値が“６１４．５”以下ならば、分類クラスはC_4、そうでなく、特徴量N_1の値が“６１４．５”より大きく“６４９．５”以下ならば、分類クラスはC_5、そうでなく…」と云う知識を表す。

【０００６】図９は、公知の決定木学習装置の構成を示したものであり、図１０は公知の決定木学習装置の処理の流れを示したものである。図９において、９１は決定木学習用データ入力部であって、当該決定木学習装置に決定木学習用データを入力するものであり、９２は決定木生成終了判定部であって、特徴量選択部９４によって選択された特徴量がどの値であるかによって、与えられたデータＳを複数の部分に分割し、それぞれの部分について図１０のＳｔ８１のステップ以降の処理を再帰的に繰り返すことで、全ての枝が終了したか否かを判断するものであり、この時、生成中の決定木９９に、選択された特徴量を付け加えることで決定木を成長させるのが決定木生成部９３である。９４は特徴量選択部であって、特徴量がどの値であるかを選択するものである。また、９５は連続値を区間に分割することによって記号的特徴量に変換する連続値特徴量記号化装置、９６は相互情報量計算部であって、与えられたデータＳの持つ情報量Ｉ（Ｓ）と、すべての特徴量A_iについて、その特徴量がどの値であるかによってデータを分割した後の各部分の情報量の和Ｅ（A_i，Ｓ）、さらにその差gain（A_i，Ｓ）＝Ｉ（Ｓ）－Ｅ（A_i，Ｓ）を計算するものであり、この差gain（A_i，Ｓ）が最も大きい特徴量を選ぶのが上述の特徴量選択部９４である。

【０００７】９７は決定木出力部であって、決定木生成終了判定部９２が決定木生成終了と判定した場合に、生成中の決定木９９を完成した決定木として出力するものであり、９８がこの決定木出力部９７の出力した決定木である。

【０００８】このような構成の従来装置は、つぎのような処理を行う。決定木学習用データ入力部９１から決定木学習用データが入力されると（図１０のＳｔ８０）、決定木生成終了判定部９２によって、分類クラスに分けられた事例データが、その分類クラスに該当するかの誤り率である分類誤り率設定値（prune rate）以上を占めるか否かが判断される（図１０のＳｔ８１）。

【０００９】もしそうであって、すべての再帰的繰り返しについて、やはり同じことが言えているならば（図１０のＳｔ８２）、決定木出力部９７が生成中の決定木９９を決定木９８として出力し（図１０のＳｔ８３）、処理を終了する。

【００１０】そうでないならば、決定木生成部９３は、特徴量選択部９４によって選択された特徴量がどの値で

3

あるかによって、与えられたデータＳを複数の部分に分割し、それぞれの部分について（図１０のＳｔ８１）以降の処理を再帰的に繰り返すことを行う。この時、生成中の決定木（９９）に選択された特徴量を付け加えることで決定木を成長させる。

【００１１】以上の処理における特徴量の選択は、つぎにようにして行う。すなわち、相互情報量計算部９６により、与えられたデータＳの持つ情報量Ｉ（Ｓ）と、すべての特徴量A_iについて、その特徴量がどの値である＊

4

＊かによってデータを分割した後の各部分の情報量の和Ｅ（A_i，Ｓ）、さらにその差gain（A_i，Ｓ）＝Ｉ（Ｓ）－Ｅ（A_i，Ｓ）を計算し（Ｓｔ８５）、差gain（A_i，Ｓ）が最も大きい特徴量を選ぶ（Ｓｔ８７）ことによって行う。ここで、分類クラスをC_0，C_1，…，C_nとし、上記与えられたデータＳ中でのクラスC_iの出現頻度をP_i（Ｓ）としたとき、

【００１２】

【数１】

$$I(S) = -\sum_{k=1}^{n} P_k(S) \log_2 P_k(S)$$

$$E(A_i, S) = \sum_{S_j \in A_i \text{によるSの分割}} \frac{|S_j|}{|S|} I(S_j)$$

である。

【００１３】なお、これについては、「参考文献２」（Quinlan, J. R., “Induction of Decision Trees”, Machine Learning Vol. 1, 1986.）に詳しい。この時、もし特徴量中に連続値をとるものがあったならば（図１０のＳｔ８４）、連続値特徴量記号化装置９５によって、連続値を区間分割することによって記号的特徴量に変換し（図１０のＳｔ８６）、記号的特徴量と同様に扱う。

【００１４】以下では、連続値特徴量記号化装置９５において用いられる公知の区間分割アルゴリズムの例について説明する。図１１は、連続値特徴量記号化装置９５において用いられる公知の区間分割アルゴリズムの例の構成を示したものであり、図１２は、連続値の区間を分割する処理（図１０のＳ８６）において用いられる公知の区間分割アルゴリズムの例における処理の流れを示したものである。

【００１５】学習データ集合／連続値特徴量入力部７１には、決定木学習用データと、区間を分割すべき連続値特徴量とが入力される（図１２のＳｔ３０）。入力された特徴量をＡとする。

【００１６】次に、学習データ分類部７２が、入力された決定木学習用データ７９を、その分類クラスによって分類する（図１２のＳｔ３１）。次に、分類結果序列化装置７３が、分類された学習用データ８０の各分類結果を、それぞれ特徴量Ａの値の平均値により順序付ける（図１２のＳｔ３２）。序列化された分類済み学習用データ８１に対し、境界値決定部７４は、隣合った分類結果w_1、w_2に対して、以下の式で決められる境界値Ｂを計算する。ここで、m_iは分類結果w_iの特徴量Ａの平均を、d_iは分類結果w_iの特徴量Ａの値の分散を表す。

【００１７】

$$B = (m_2 d_1 + m_1 d_2) / (d_1 + d_2)$$

こうして、すべての分類結果に対し境界値を割り当てたなら、再分割必要性判定部７５が、再度分割すべき分類結果があるかどうかを判定する。これは、分類結果w_iに対し、w_i中のデータで、w_iに割り当てられた境界値内に特徴量Ａの値が属さないものの割合が、予め決められた割合（cluster rate）より大きいかどうかを判定することにより行う（図１２のＳｔ３４）。

【００１８】その結果、もし大きいなら、例題分割部７６が、その分類結果を２つに分割する（図１２のＳｔ３５）。分割は、例えば、後述するk-means アルゴリズムを、ｋ＝２として用いる。

【００１９】ここで、一つでも分割された分類結果があったならば、図１２のＳｔ３２以降を繰り返す（図１２のＳｔ３６）。もしなかったならば、区間記号化装置７７が、分割された各区間に記号を付与する（図１２のＳｔ３７）。そして、記号化特徴量出力部７８が、特徴量Ａを記号化した特徴量を出力する（図１２のＳ３８）。

【００２０】なお、これについては、「参考文献１」（荒木大，小島昌一，“決定木学習における数値データの区間分割”，第５回人口知能学会大会論文集，1991.）に詳しい。

【００２１】図８は、図７に示す事例データから、上記定められた割合の設定値（prune rate）を７０％、与えられたクラスタ誤り率のレート（cluster rate）を４０％として上記決定木学習装置を用いて導いた決定木である。

【００２２】

【発明が解決しようとする課題】以上述べた従来の技術では、以下のような問題点がある。従来の決定木学習装置においては、事例データ中に分類クラスが与えられることを前提としており、しかもそれは離散的な値をとるものでなければならない。従って、ある特徴量にしたがって分類を行うような決定木を求めたいが、その特徴量が連続値をとる場合には、事前にオペレータ等が区間分割等を行って、離散な分類クラスを定義してやる必要があった。

【００２３】例えば、図３に示すような事例データが与

えられ、分類対象の特徴量N0 およびN2 の値に従って分類を行う決定木を求めたいとするならば、たとえば特徴量N0 に対して　｛６２０未満、６２０以上６４０未満、６４０以上６６０未満、６６０以上６８０未満、６８０以上｝　と云うような区間分割を行い、この区間分割によって得られたそれぞれの区間に、N2 の値であるAの場合とBの場合とで例えば、　｛C1 , C2 , C3 , C4 , C5 ｝　と云うような具合に分類クラスの名前を与えた後に、事例データを決定木学習装置に入力する必要があった。図７に示す事例データは、図３に示すものに対してこのような操作を人手によって行い、得られたものを示したものである。

【００２４】しかし、人手によるこのような作業は、量が少ない場合は良いが、量が膨大になったり、分類が複雑になると手に負えなくなる。従って、この作業を自動化する必要がある。

【００２５】本発明の目的とするところは、連続値をとる特徴量によって分類を行うような決定木を求めたい場合に、その特徴量に対して適切な分類クラスを自動的に設定し、連続値をとる特徴量による分類のための決定木を自動的に導くことが出来るようにした決定木学習装置を提供することにある。

【００２６】

【課題を解決するための手段】上記目的を達成するため、本発明はつぎのように構成する。すなわち、事例の特徴を表す第１の特徴量と、分類用の属性を示す第２の特徴量との少なくとも２種の特徴量を有する事例データを対象とし、複数の事例データの集合に対し、前記複数種の特徴量に応じてその事例データ集合を、予め定めた分類クラスのうちの対応する分類クラスに分類するための規則を得る決定木学習生成装置として、与えられた複数の事例データを順次取り込み、その第１の特徴量の最大値と最小値との差を属性別に求め、この差の分割後の値が前記第１の特徴量を分類するために予め定めた分類の粒度の値以下となるまで、前記差と当該粒度の値とにより定まるクラスタ数分に分割して分類クラスとすると共に、各隣接する分類クラスの特徴値の境界値範囲を前記分割の際の各分類クラスの特徴値のうちの最も近い値の中間値を以て定めることにより決定木を生成する手段、分類クラスそれぞれの分割クラスのクラス分け誤り率が予め定めた設定値に達すると決定木生成を終了する決定木生成終了制御手段とを設けて構成した。

【００２７】

【作用】このような構成において、複数の事例データを与えると、決定木生成手段は当該与えられた複数の事例データを順次取り込み、その第１の特徴量の最大値と最小値との差を属性別に求め、この差の分割後の値が前記第１の特徴値（特徴量）を分類するために予め定めた分類の粒度の値以下となるまで、予め設定されたクラスタ数分に分割して分類クラスとすると共に、各隣接する分類クラスの特徴値の境界値範囲を前記分割の際の各分類クラスの特徴値のうちの最も近い値の中間値を以て定めることにより決定木を生成する。そして、決定木生成終了制御手段は、分類クラスそれぞれの分割クラスのクラス分け誤り率が予め定めた設定値に達するか、またはデータ集合の分類クラスについて、定義されるべき特徴量の分散が設定精度に達すると決定木生成を終了させる。

【００２８】本発明は、事例の特徴を表す第１の特徴量と、分類用の属性を示す第２の特徴量との少なくとも２種の特徴量を有する事例データを対象とし、複数の事例データの集合に対し、前記複数種の特徴量に応じてその事例データ集合を、予め定めた分類クラスのうちの対応する分類クラスに分類するための規則である決定木を、事例データ集合を与えることで生成する決定木学習装置であり、連続値をとる特徴量と属性によってデータの分類を行う場合の決定木導出における各特徴量選択の段階でクラス分け（クラスタリング）を行い、適切な区間をこの特徴量に対して設定する。クラスタリング・アルゴリズムに対するパラメータであるクラスタ数は、ユーザが必要とする分類の精度（分類の粒度）を与えることによって自動的に決定する。決定木導出は、データ集合の分類クラスについて、定義されるべき特徴量の分散が設定精度より小さくなるか、あるいは、その集合の、クラスタリングによって与えられた分類クラスの中でのクラス分け誤り率が設定レート（prune rate）に収まるか、またはデータ集合の分類クラスについて、定義されるべき特徴量の分散が設定精度に収まると終了させるようにした。

【００２９】このように、連続値を属性に対応して分類する決定木を作成しようとする場合、分類の精度（分類の粒度）と、特徴値と分類すべき属性の値からなる学習用の複数のサンプルデータを与えることによって当該サンプルデータから該サンプルデータの属性別最大差を前記分類の粒度に従った分類クラスにクラスタリングし、その分類クラスの特徴値境界値範囲を前記分割の際の各分類クラスの特徴値のうちの最も近い値の中間値を以て定めることにより決定木を生成することから分類クラスを自動的に定義できるので、連続値属性に対する分類木の作成が完全に自動化される。

【００３０】従って本発明によれば、連続値をとる特徴量によって分類を行うような決定木を求めたい場合に、その特徴量に対して適切な分類クラスを自動的に設定して、連続値をとる特徴量による分類のための決定木を自動的に導くことが出来るようになる決定木学習装置を提供することができる。

【００３１】

【実施例】以下、本発明の実施例を図面に基づいて説明する。図１は、本発明の構成を示すものであり、図２は本発明全体の処理の流れを示すものである。図１において、５１は決定木学習用データ入力部、５２は決定木生

成終了判定部、５３は決定木生成部、５４は特徴量選択部、５５は連続特徴量記号化装置、５６は相互情報量計算部、５７は決定木出力部、５８は決定木、５９は生成中の決定木、６０はクラスタリング部、６１はパラメータ入力部である。

【００３２】これらのうち、決定木学習用データ入力部５１はいくつかの特徴量からなる事例のデータを入力するものであり、パラメータ入力部６１は分類の対象にする属性名やその分類に必要とされる精度（分類の粒度（prec））、決定木生成終了判定に用いる同一の分類クラスに分類されたデータの分類誤り率を示す分類誤り率設定値（prune rate）、そして、与えられたクラスタ誤り率のレート（clusterrate）などのパラメータ・データを入力するためのものである。

【００３３】また、決定木生成終了判定部５２は、与えられた訓練例の属性Ｎに属する枝の取り得る値の最大値と最小値との差（diff）が、分類の粒度（prec）以下であるかどうかを判断し、差（diff）が分類の粒度prec以下であって、すべての再帰的繰り返しについて同じことが言えている場合に、全ての再帰の分枝が終了のときは終了と判断し、全ての再帰の分枝が終了していなければその枝は終りと判定してつぎの枝の処理を実行させる指示を行う機能を有する。

【００３４】決定木生成部５３は、特徴量選択部５４によって選択された特徴量がどの値であるかによって、与えられた訓練例を複数の部分に分割し、それぞれの部分について（図２のＳｔ２１）以降の処理を再帰的に繰り返すことを実施するものであり、連続特徴量記号化装置５５は相互情報量計算部５６が求める特徴量中に連続値をとるものがあった時、パラメータ・データである与えられたクラスタ誤り率のレート（cluster rate）を用いてこのレートを満たすことができるならば、前記連続値を区間分割すると云った処理を行うものである。連続特徴量記号化装置５５の構成は図１１に示した公知の構成で良く、処理の流れは図１２に示した公知のアルゴリズムを利用する。

【００３５】また、相互情報量計算部５６は、与えられた訓練例集合Ｓの持つ情報量Ｉ（Ｓ）と、全ての特徴量Ａｉについて、その特徴量がどの値であるかによってデータを分割した後の各部分の情報量の和Ｅ（Ａｉ，Ｓ）、さらにその差gain（Ａｉ，Ｓ）＝Ｉ（Ｓ）－Ｅ（Ａｉ，Ｓ）を計算するものであり、特徴量選択部５４はこの結果から、上記差gain（Ａｉ，Ｓ）が最も大きい特徴量を選ぶと云った処理を行うものである。

【００３６】決定木出力部５７は決定木生成終了判定部５２が全ての再帰の分枝終了と判断した際に、決定木生成部５３が生成して生成中の決定木データ５９として保存した当該決定木データ５９を最終的な決定木データ５８として出力するものである。クラスタリング部６０は訓練例の属性Ｎの値の集合に対し、「diff／prec」個のクラスタに分割するクラスタリングを行うものである。

【００３７】つぎに上記構成の本装置の作用を図２のフローチャートを参照して説明する。いくつかの特徴量からなる事例（以下、訓練例と呼ぶ）が紙、または磁気テープなどに表形式に記録されているとする。例えば図３である。

【００３８】図３は、Ｎ０、Ｎ１、Ｎ２と云う３つの特徴量を持った訓練例である。この訓練例のデータは、キーボード入力、ネットワークによるオンライン入力、情報伝達の媒体である磁気テープからの読取り入力等のかたちで決定木学習用データ入力部５１より入力される（図２のＳｔ２０）。また、パラメータ入力部６１からは、分類の対象にする属性名（与えられた事例データの、どの特徴量に基づいて分類を行うかを示すもの）、その分類に必要とされる精度（分類の粒度（prec））、前述の、決定木生成終了判定のために用いる同一の分類クラスに分類されたデータの分類誤り率の割合を示す分類誤り率設定値（prune rate）、連続値特徴量記号化装置５５で用いられるクラスタ誤り率のレート（cluster rate）などのパラメータ・データが入力される。

【００３９】決定木学習データ入力部５１から訓練例が入力されると（図２のＳｔ２０）、決定木生成終了判定部５２によって、与えられた訓練例の属性Ｎの値の最大値と最小値との差（diff）が分類の粒度であるprec以下であるか否かが判断される（図２のＳｔ３０）。その結果、差（diff）が分類の粒度prec以下であって、すべての再帰的繰り返しについてやはり同じことが言えているならば、全ての再帰の分枝が終了と判断し（図２のＳｔ２２）、これによって決定木出力部５７が決定木のデータ５８を出力し（図２のＳｔ２３）、決定木を求める処理を終了する。

【００４０】そうでないときはクラスタリング部６０でのクラスタリング処理に入る。クラスタリング部６０では、訓練例の属性Ｎの値の集合に対し、「diff／prec」個のクラスタに分割すると云ったクラスタリングの処理を行う（図２のＳｔ２９）。これには、例えば公知のk-means アルゴリズムを用いる。

【００４１】すなわち、クラスタリング部６０はクラスタリングの処理をつぎのようにして行う。今、各クラスタを分類のためのクラス（以下分類クラス）と考える。この時、各クラスタの境界値は、そのクラスタの最小値と値の小さい側に隣合ったクラスタの最大値との中間値、およびそのクラスタの最大値と値の大きい側に隣合ったクラスタの最小値との中間値とする。最小（最大）のクラスタの小さい（大きい）側の境界値は設定されないものとする。例えば、

｛６６７｝，｛６７４，６８１｝，｛６８８｝

と云うクラスタが出来た時、｛６６７｝のクラスタは「無限小から“６７０．５”以下」と云う名前の分類クラスと考え、｛６７４，６８１｝のクラスタは「“６７

9

０．５”より大きく、“６８４．５”以下」と云う名前の分類クラスと考えるのである。

【００４２】クラスタリング部６０によって分類クラスが定義された後は、前述した公知の決定木学習装置と同様に、特徴量を選択することによって決定木を生成する。すなわち、これはつぎのようにして行う。

【００４３】クラスタリング部６０による分類クラスの定義が終了すると、決定木生成終了判定部５２での処理に入り、まず、決定木生成終了判定部５２において、分類クラスに分類された事例データの分類誤り率が、同一の分類クラスに分類されたデータの分類誤り率設定値（prune rate）以上を占めるかどうかが判断される（図２のＳｔ２１）。その結果、もし分類誤り率設定値（prune rate）以上を占めていて、すべての再帰的繰り返しについてやはり同じことが言えているならば（図２のＳｔ２２）、決定木出力部５７が決定木（５８）を出力し（Ｓｔ２３）、決定木を求める処理を終了する。

【００４４】図２のステップＳｔ２１の判定の結果、もし分類誤り率設定値（prune rate）以上を占めていないならば、決定木生成部５３での処理に移る。そして、決定木生成部５３では、特徴量選択部５４によって選択された特徴量がどの値であるかによって、与えられた訓練例を分割してそれぞれの部分について（図２のＳｔ２１）以降の処理を再帰的に繰り返すことを実施し、これによって、当該訓練例の取り得る枝を複数に分岐し、それぞれの部分についての特徴量を求める。

【００４５】この時、生成中の決定木５９の枝に対して、上述の選択された特徴量を付け加えることで決定木生成部５３は決定木を成長させる。以上の処理における特徴量の選択は、相互情報量計算部５６の計算結果をもとに特徴量選択部５４により行われる。すなわち、相互情報量計算部５６では与えられた訓練例集合Ｓの持つ情報量Ｉ（Ｓ）と、全ての特徴量Ａｉについて、その特徴量がどの値であるかによってデータを分割した後の各部分の情報量の和Ｅ（Ａｉ，Ｓ）、さらにその差gain（Ａｉ，Ｓ）＝Ｉ（Ｓ）－Ｅ（Ａｉ，Ｓ）を計算するので（図２のＳｔ２５）、特徴量選択部５４はこの相互情報量計算部５６が求めたもののうち、上記差gain（Ａｉ，Ｓ）が最も大きい特徴量を選ぶ（図２のＳｔ２７）。これによって特徴量の選択が成される。

【００４６】この時、もし特徴量中に連続値をとるものがあったならば（図２のＳｔ２４）、連続値特徴量記号化装置５５によって、連続値を区間に分割することによって記号的特徴量と同様に扱う。この分割処理は図１２のフローチャートに従う。

【００４７】以下では、クラスタリング部６０において用いられる公知のクラスタリングアルゴリズムの例について説明する。図４は、クラスタリング部６０において用いられる公知のクラスタリング・アルゴリズム例の構成を示したものであり、図５は、分類対象特徴量のクラスタリング（図２のＳｔ２９）において用いられる公知のクラスタリング・アルゴリズム例における処理の流れを示したものである。

10

【００４８】クラスタリング部６０によるクラスタリング処理は図４に示すように、クラスタリングされるべきデータと、パラメータ（ｋ）とを入力するためにあるデータ／ｋ入力部１１から、まず、クラスタリングされるべきデータと、いくつのクラスタリングに分けるかと云うパラメータ（ｋ）が入力されることにより開始される（図５のＳｔ４１）。データ／ｋ入力部１１から入力されたこれらのデータは、データ記憶部１２に保持される。

【００４９】このデータ記憶部１２に保持されたデータは、クラスタリング実行部１７と平均値初期化部１３に与えられる。すると、平均値初期化部１３はクラスタの平均を、データ記憶部１２に保存されたデータの先頭から、相異なるｋ個を取ってその平均値を求め、各クラスタの平均値保持部１５に更新記憶させることによって各クラスタの平均値の初期化をする（図５のＳｔ４２）。

【００５０】一方、クラスタリング実行部１７は、データ記憶部１２に記憶されている各データを、それと最も近い平均値を持つクラスタへと分類する（図５のＳｔ４３（クラスタリング処理））。

【００５１】そして、その結果はクラスタリング結果としてクラスタリング結果保存部１８保存する。ついで各クラスタの平均値保持部１５に保持されている平均値ＤAV-NEWを前回の平均値ＤAV-OLDとして前回の各クラスタ平均値保持部１４へとコピーした後、クラスタリング結果保存部１８に対して、平均値計算部２０が各クラスタの平均値を計算し（図５のＳｔ４４）、この計算により得られた平均値を新しい平均値ＤAV-NEWとして各クラスタの平均値保持部１５に保存する。

【００５２】ここで、クラスタ平均比較部１６が、各クラスタの平均値保持部１５に保存された新しい平均値ＤAV-NEWと、前回の各クラスタ平均値保持部１４に保持されている前回の平均値ＤAV-OLDとを比較して両者が等しいか否かを判断する（図５のＳｔ４５）。

【００５３】クラスタ平均比較部１６による当該判断の結果、両者が等しければ、出力部１９はクラスタリング結果保存部１８からクラスタリング結果を出力して（図５のＳｔ４６）終了する。

【００５４】ステップＳｔ４６の比較の結果、両者が等しくなければステップＳｔ４３以下の処理を繰り返す。以上がクラスタリング部６０におけるクラスタリング処理操作の詳細である。

【００５５】図６は、図３の学習用データに対して、分類対象特徴量をＮ０、分類の粒度（prec）を“１０”、与えられたクラスタ誤り率のレート（cluster rate）を“４０％”、分類誤り率設定値（prune rate）を“９９％”として本発明を用いて生成した決定木である。

11

【００５６】図６の決定木は、「特徴量N1 が“６２１．５”以下であり、特徴量N2 がＡであるなら、特徴量N0 の値は“６７０．５”から“６８４．５”の間である。そうでなく（N2 がＢで）、N1 が以下なら、N0 の値は“６５６．５”から“６７０．５”の間である。そうでなく…」と云う知識を表す。

【００５７】従来の方法においては、決定木の“葉”に現れる値（の範囲）は、学習システムの外側でユーザが与える必要があったが、本発明によれば、決定木の“葉”に現れる値は、与えられた精度に対して適切な値（適切な値の範囲）となるように自動的に決定できる。

【００５８】要するに、本発明は、数値または記号で表現された幾つかの特徴量と、それに対してオペレータ等が事前にデータを吟味して与える分類結果を事例とし、多数の事例を分類装置の例題として与えることによって、特徴量と分類結果の間の一般法則を見付け出し、新たな事例が与えられた時に、その事例の持つ特徴量を評価して分類結果を出力すると云った処理をする学習装置において、事例の特徴を表す第１の特徴量と、分類用の属性を示す第２の特徴量との少なくとも２種の特徴量を有する事例データを対象とし、複数の事例データの集合に対し、前記複数種の特徴量に応じてその事例データ集合を、予め定めた分類クラスのうちの対応する分類クラスに分類するための規則を得る決定木学習装置として、与えられた複数の事例データを順次取り込み、その第１の特徴量の最大値と最小値との差を属性別に求め、この差の分割後の値が前記第１の特徴量を分類するために予め定めた分類の粒度の値以下となるまで、前記差と当該粒度の値とにより定まるクラスタ数分に分割して分類クラスとすると共に、各隣接する分類クラスの特徴値の境界値範囲を前記分割の際の各分類クラスの特徴値のうちの最も近い値の中間値を以て定めることにより決定木を生成する手段、分類クラスそれぞれの分割クラスのクラス分け誤り率が予め定めた設定値に達するか、またはデータ集合の分類クラスについて、定義されるべき特徴量の分散が設定精度に達すると決定木生成を終了する決定木生成終了制御手段とを設けて構成したものである。

【００５９】そして、このような構成において、複数の事例データを与えると、決定木生成手段は当該与えられた複数の事例データを順次取り込み、その第１の特徴量の最大値と最小値との差を属性別に求め、この差の分割後の値が前記第１の特徴量を分類するために予め定めた分類の粒度の値以下となるまで、前記差と当該粒度の値とにより定まるクラスタ数分に分割して分類クラスとすると共に、各隣接する分類クラスの特徴値の境界値範囲を前記分割の際の各分類クラスの特徴値のうちの最も近い値の中間値を以て定めることにより決定木を生成し、そして、決定木生成終了制御手段は、分類クラスそれぞれの分割クラスのクラス分け誤り率が予め定めた設定値に達するか、またはデータ集合の分類クラスについて、定義されるべき特徴量の分散が設定精度に達すると決定木生成を終了させるものである。

12

【００６０】本発明は、事例の特徴を表す第１の特徴量と、分類用の属性を示す第２の特徴量との少なくとも２種の特徴量を有する事例データを対象とし、複数の事例データの集合に対し、前記複数種の特徴量に応じてその事例データ集合を、予め定めた分類クラスのうちの対応する分類クラスに分類するための規則である決定木を、事例データ集合を与えることで生成する決定木学習装置であり、連続値をとる特徴量と属性によってデータの分類を行う場合の決定木導出における各特徴量選択の段階でクラス分け（クラスタリング）を行い、適切な区間をこの特徴量に対して設定する。そして、クラスタリングに必要なパラメータであるクラスタ数は、ユーザが必要とする分類の精度（分類の粒度）を与えることによって自動的に決定するようにしており、また、決定木導出は、データ集合の分類クラスについて、定義されるべき特徴量の分散が設定精度より小さくなるか、あるいは、その集合の、クラスタリングによって与えられた分類クラスの中でのクラス分け誤り率が設定レート（prune rate）に収まれば終了させるようにした。

【００６１】そしてこのように、連続値を属性に対応して分類する決定木を作成しようとする場合、分類の精度（分類の粒度）と、特徴値と分類すべき属性の値からなる学習用の複数のサンプルデータを与えることによって当該サンプルデータから該サンプルデータの属性別最大差を前記分類の粒度に従った分類クラスにクラスタリングし、その分類クラスの特徴値境界値範囲を前記分割の際の各分類クラスの特徴値のうちの最も近い値の中間値を以て定めることにより決定木を生成する方式を採用したことにより、分類クラスを自動的に定義できるので、連続値属性に対する分類木の作成を完全に自動化することができるようになる。

【００６２】従って本発明によれば、連続値をとる特徴量によって分類を行うような決定木を求めたい場合に、その特徴量に対して適切な分類クラスを自動的に設定して、連続値をとる特徴量による分類のための決定木を自動的に導くことが出来るようになる決定木学習生成装置が得られる。なお、本発明は上述した実施例に限定するものではなく、その要旨を変更しない範囲内で適宜変形して実施し得るものである。

【００６３】

【発明の効果】以上詳述したように本発明によれば、連続値属性に対する分類木を作ろうとする場合、クラスタリングによって分類クラスが自動的に定義されるので、分類木の作成が完全に自動化されるなど、連続値をとる特徴量によって分類を行うような決定木を求めたい場合に、その特徴量に対して適切な分類クラスを自動的設定して、連続値をとる特徴量による分類のための決定木を自動的に導くことが出来るようになる決定木学習生成装

置を提供できる。

【図面の簡単な説明】

【図１】本発明の実施例を説明するための図であって、本発明の一実施例の構成図。

【図２】本発明の実施例を説明するための図であって、本発明の一実施例における装置の動きを示す流れ図。

【図３】本発明の実施例を説明するための図であって、本発明の実施例で入力として用いる訓練例のデータを示す図。

【図４】本発明の実施例を説明するための図であって、本発明の実施例におけるクラスタリング部６０の詳しい構成図。

【図５】本発明の実施例を説明するための図であって、本発明の一実施例におけるクラスタリング部６０の装置の詳しい動きを示す流れ図。

【図６】本発明の実施例を説明するための図であって、図３に示すデータと、所定の学習用パラメータに対し、本発明の一実施例である装置が生成した決定木を示す図。

【図７】従来技術を説明するための図であって、公知の決定木学習装置が扱える事例データの例を示す図。

【図８】従来技術を説明するための図であって、図７の事例データに対して公知の決定木学習装置が導いた決定*木を示す図。

【図９】従来技術を説明するための図であって、公知の決定木学習装置の構成を示すブロック。

【図１０】従来技術を説明するための図であって、公知の決定木学習装置の処理の動きを表す流れ図。

【図１１】従来および本発明の実施例において使用される連続値特徴量記号化装置９５，５５の詳しい構成図。

【図１２】従来および本発明の実施例において使用される連続値特徴量記号化装置９５，５５の装置の詳しい動きを示す流れ図。

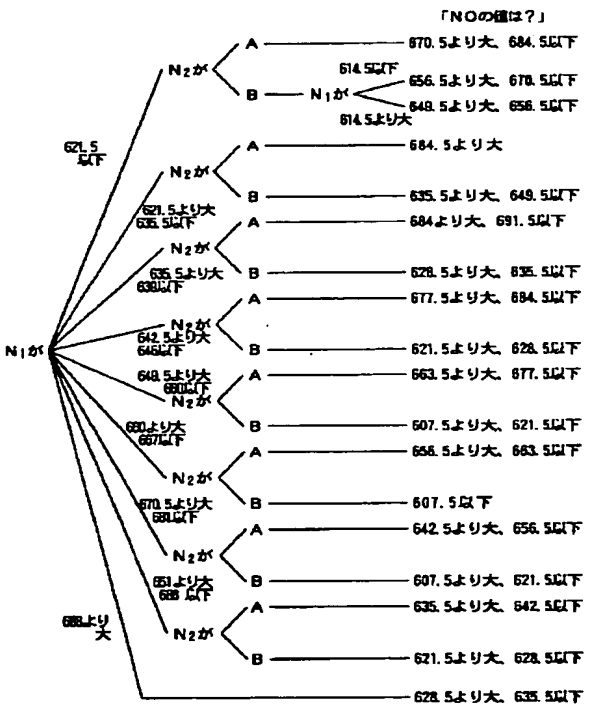
【符号の説明】

５１…決定木学習用データ入力部
５２…決定木生成終了判定部
５３…決定木生成部
５４…特徴量選択部
５５…連続特徴量記号化装置
５６…相互情報量計算部
５７…決定木出力部
５８…決定木
５９…生成中の決定木
６０…クラスタリング部
６１…パラメータ入力部

【図３】

事例番号	N0	N1	N2
1	688	639	A
2	681	646	A
3	674	653	A
4	667	660	A
5	660	667	A
6	653	674	A
7	646	681	A
8	639	688	A
9	632	695	A
10	625	688	B
11	618	681	B
12	611	674	B
13	604	667	B
14	611	660	B
15	618	653	B
16	625	646	B
17	632	639	B
18	639	632	B
19	646	625	B
20	653	618	B
21	660	611	B
22	667	604	B
23	674	611	A
24	681	618	A
25	688	625	A
26	695	632	A
27	688	639	A
28	681	646	A

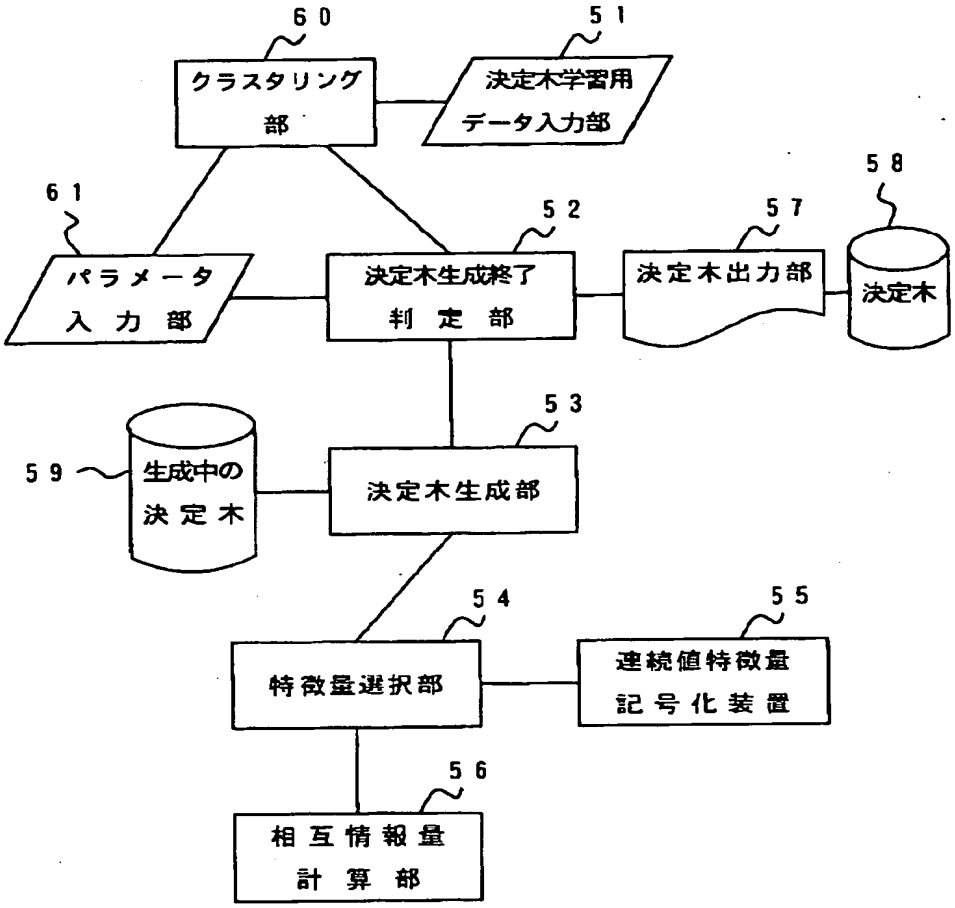
【図６】



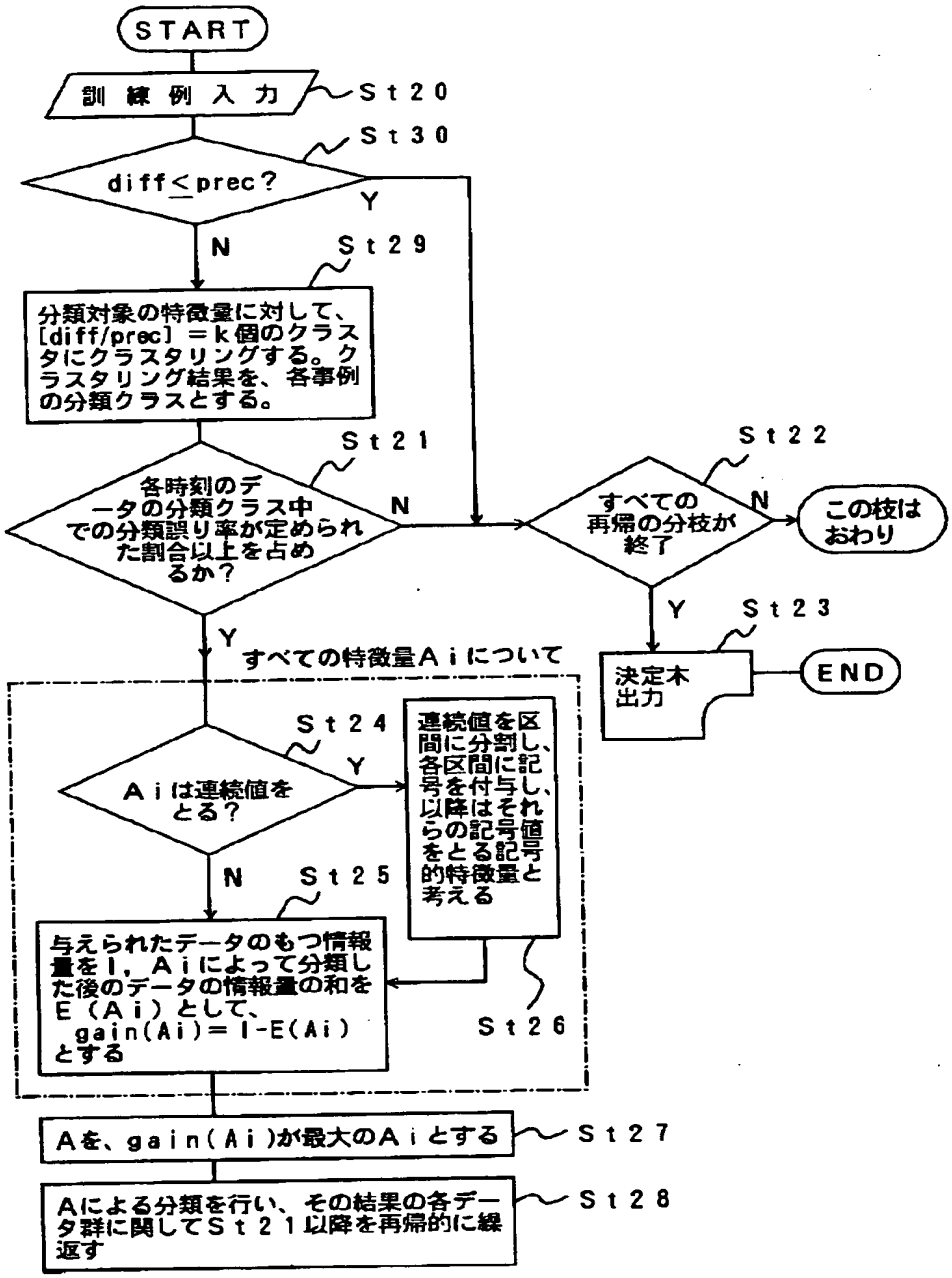
【図７】

事例番号	分類クラス	N1	N2
1	C8	639	A
2	C8	646	A
3	C4	653	A
4	C4	660	A
5	C4	667	A
6	C3	674	A
7	C3	681	A
8	C2	688	A
9	C2	695	A
10	C2	688	B
11	C1	681	B
12	C1	674	B
13	C1	667	B
14	C1	660	B
15	C1	653	B
16	C2	646	B
17	C2	639	B
18	C2	632	B
19	C3	625	B
20	C3	618	B
21	C4	611	B
22	C4	604	B
23	C4	611	A
24	C5	618	A
25	C5	625	A
26	C5	632	A
27	C5	639	A
28	C5	646	A

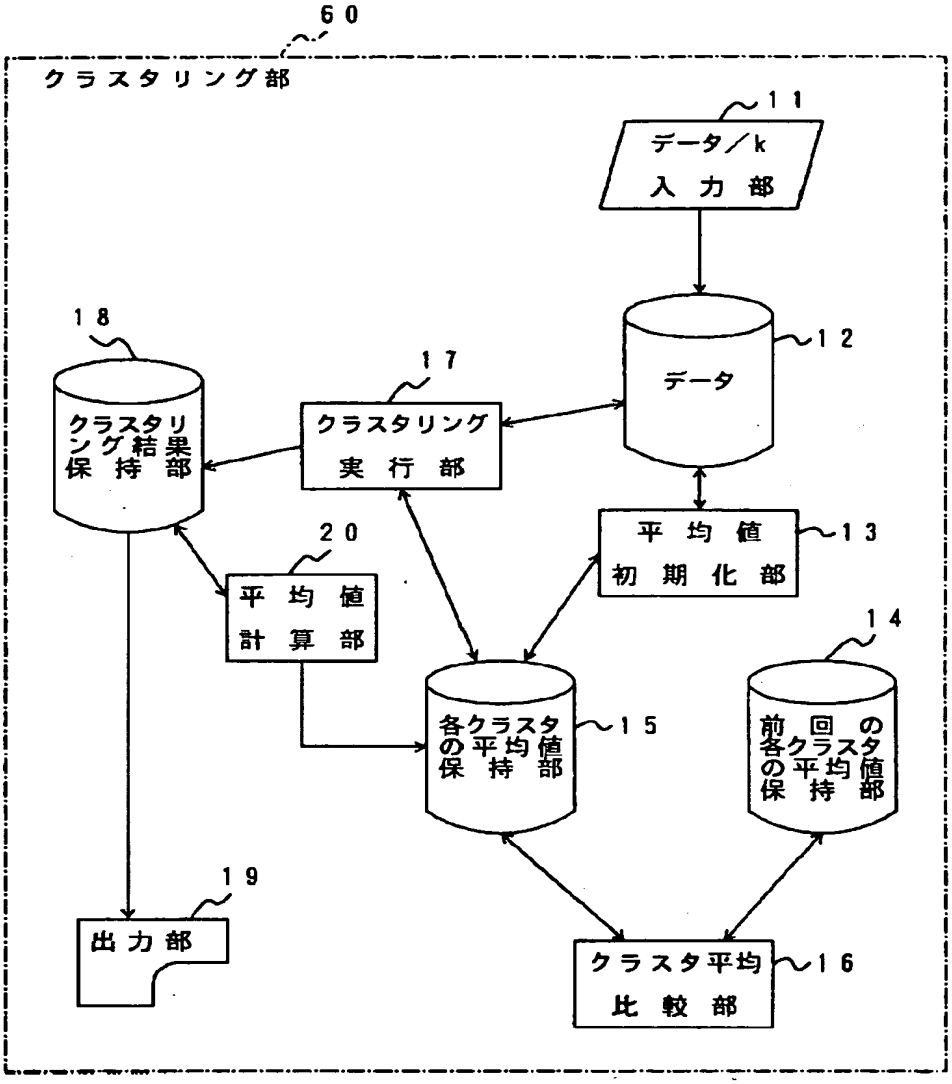
【図１】


６０
クラスタリング部
５１
決定木学習用データ入力部
６１
パラメータ入力部
５２
決定木生成終了判定部
５７
決定木出力部
５８
決定木
５９
生成中の決定木
５３
決定木生成部
５４
特徴量選択部
５５
連続値特徴量記号化装置
５６
相互情報量計算部

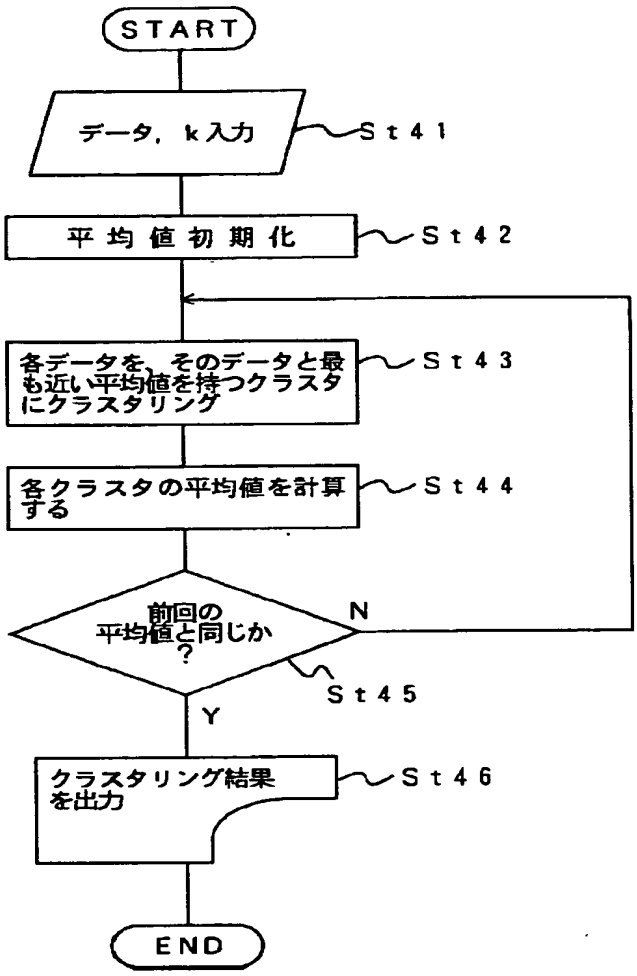
【図２】


START
訓練例入力
St20
St30
diff≦prec?
Y
N
St29
分類対象の特徴量に対して、[diff/prec]＝k個のクラスタにクラスタリングする。クラスタリング結果を、各事例の分類クラスとする。
St21
各時刻のデータの分類クラス中での分類誤り率が定められた割合以上を占めるか?
N
St22
すべての再帰の分枝が終了
N
この枝はおわり
Y
St23
決定木出力
END
Y
すべての特徴量Aiについて
St24
Aiは連続値をとる?
Y
連続値を区間に分割し、各区間に記号を付与し、以降はそれらの記号値をとる記号的特徴量と考える
St26
N
St25
与えられたデータのもつ情報量をI，Aiによって分類した後のデータの情報量の和をE(Ai)として、gain(Ai)＝I-E(Ai)とする
Aを、gain(Ai)が最大のAiとする
St27
Aによる分類を行い、その結果の各データ群に関してSt21以降を再帰的に繰返す
St28

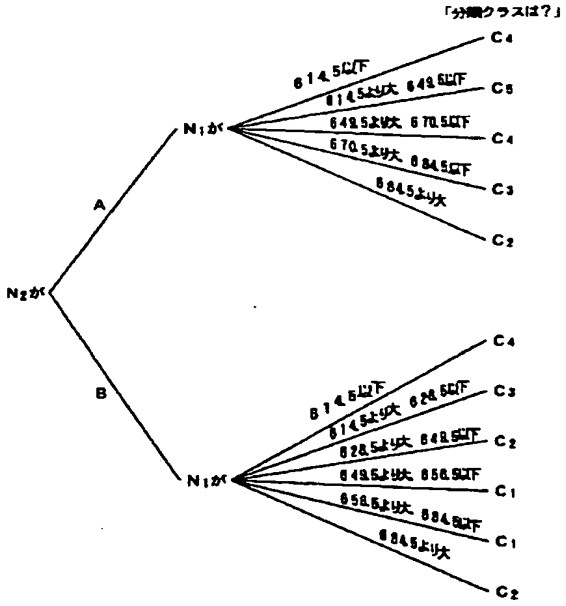
【図４】


６０
クラスタリング部
１１
データ／ｋ
入力部
１８
クラスタリング結果保持部
１７
クラスタリング実行部
データ
１２
平均値初期化部
１３
２０
平均値計算部
１４
各クラスタの平均値保持部
１５
前回の各クラスタの平均値保持部
１９
出力部
クラスタ平均比較部
１６

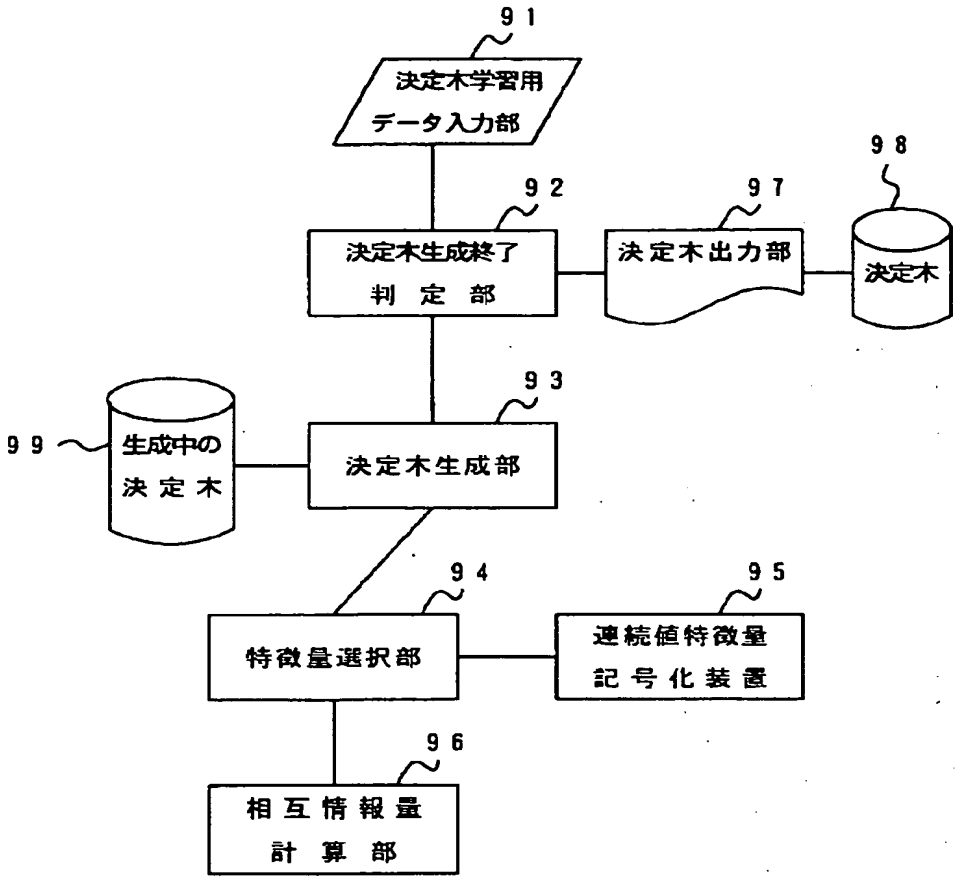
【図５】


START
データ，ｋ入力
Ｓｔ４１
平均値初期化
Ｓｔ４２
各データを、そのデータと最も近い平均値を持つクラスタにクラスタリング
Ｓｔ４３
各クラスタの平均値を計算する
Ｓｔ４４
前回の平均値と同じか？
Ｓｔ４５
N
Y
クラスタリング結果を出力
Ｓｔ４６
END

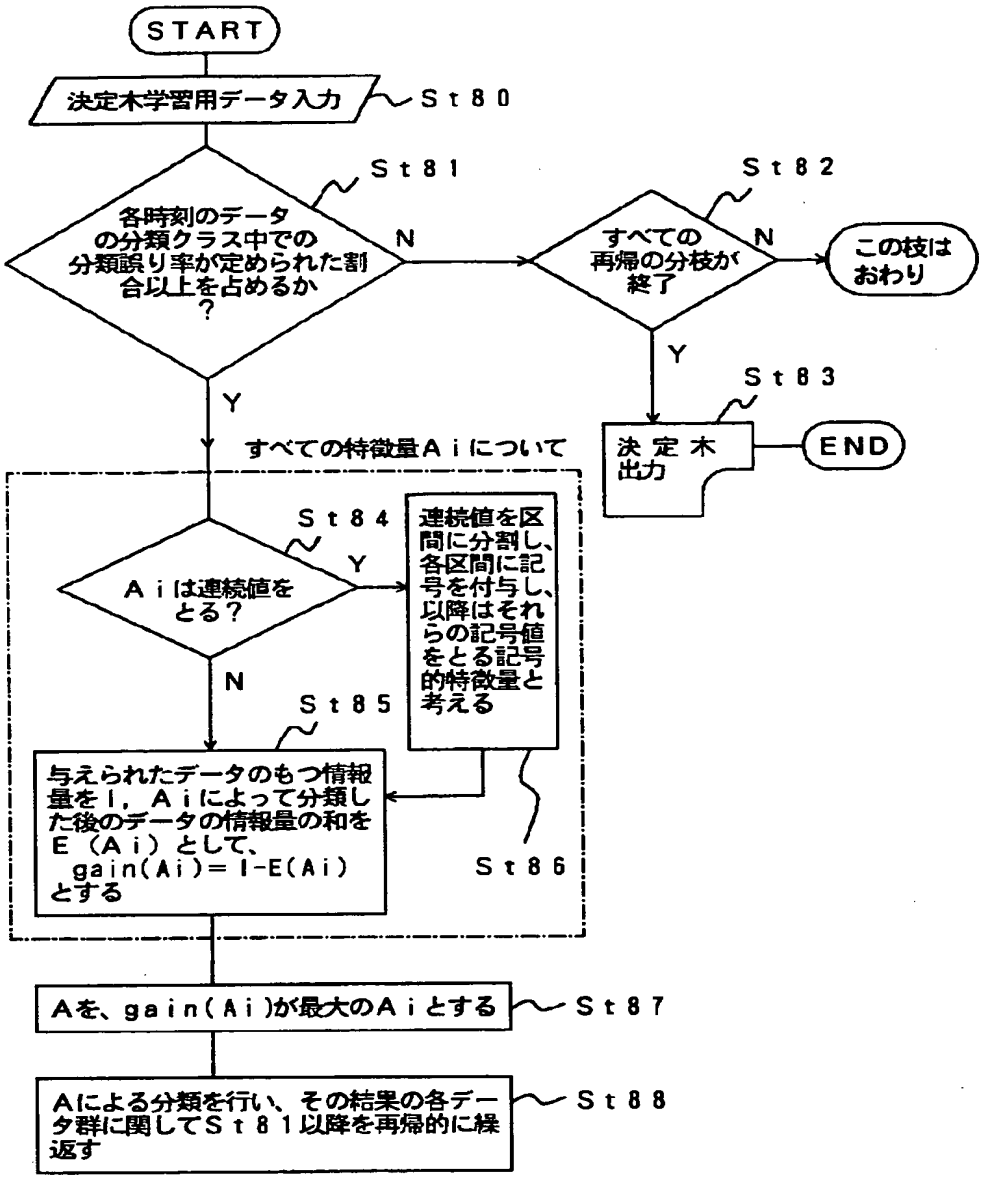
【図８】


「分類クラスは？」
N₂が
A
B
N₁が
N₁が
C₄
C₅
C₄
C₃
C₂
C₄
C₃
C₂
C₁
C₁
C₂

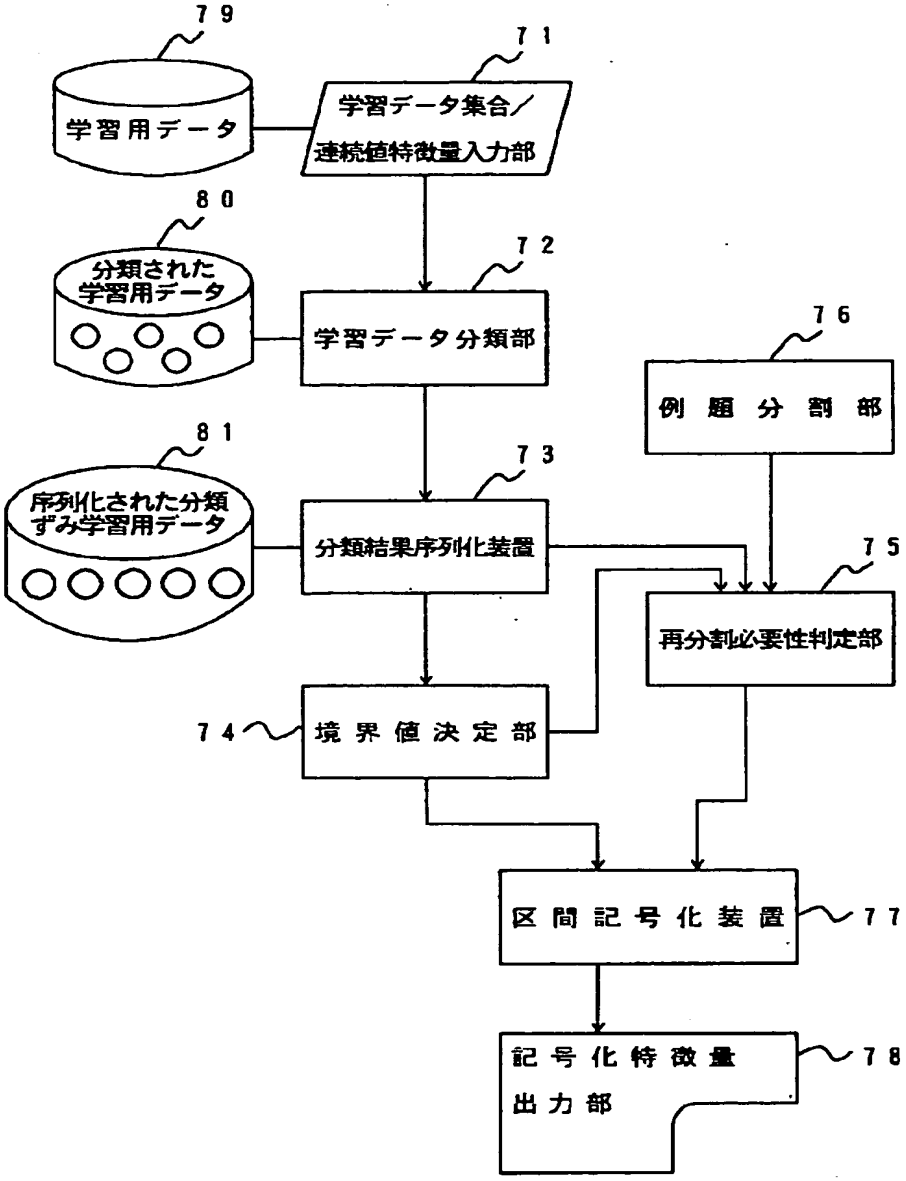
【図９】


９１
決定木学習用
データ入力部
９２
決定木生成終了
判　定　部
９７
決定木出力部
９８
決定木
９３
９９
生成中の
決　定　木
決定木生成部
９４
９５
特徴量選択部
連続値特徴量
記号化装置
９６
相互情報量
計　算　部

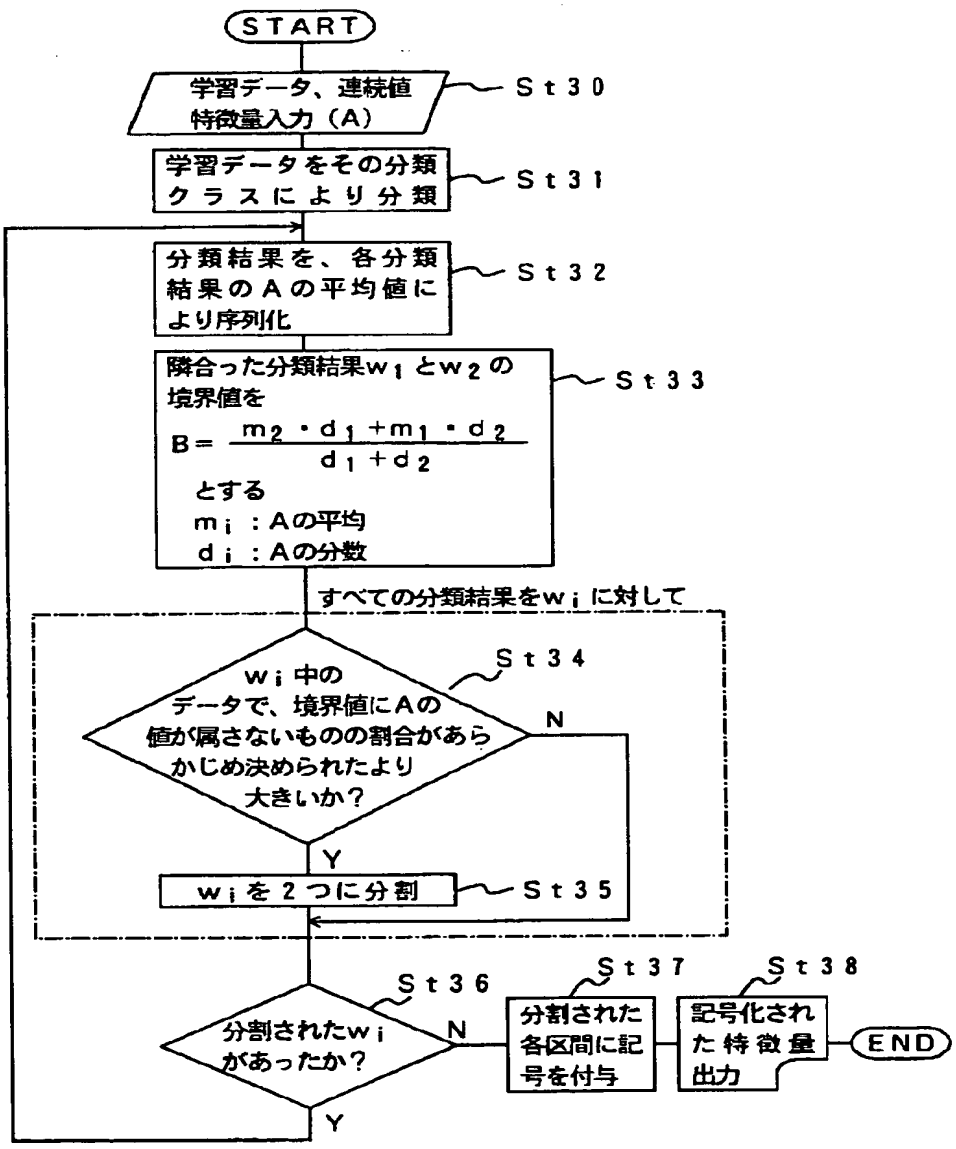
【図１０】


START
決定木学習用データ入力
St80
St81
各時刻のデータの分類クラス中での分類誤り率が定められた割合以上を占めるか？
N
St82
すべての再帰の分枝が終了
N
この枝はおわり
Y
St83
決定木出力
END
Y
すべての特徴量Aiについて
St84
Aiは連続値をとる？
Y
連続値を区間に分割し、各区間に記号を付与し、以降はそれらの記号値をとる記号的特徴量と考える
N
St85
与えられたデータのもつ情報量をI，Aiによって分類した後のデータの情報量の和をE（Ai）として、
gain(Ai)＝I-E(Ai)
とする
St86
Aを、gain(Ai)が最大のAiとする
St87
Aによる分類を行い、その結果の各データ群に関してSt81以降を再帰的に繰返す
St88

【図１１】


７９
７１
学習用データ
学習データ集合／
連続値特徴量入力部
８０
７２
分類された
学習用データ
学習データ分類部
７６
例題分割部
８１
７３
序列化された分類
ずみ学習用データ
分類結果序列化装置
７５
再分割必要性判定部
７４
境界値決定部
区間記号化装置
７７
記号化特徴量
出力部
７８

【図１２】


START
学習データ、連続値特徴量入力（A）
St30
学習データをその分類クラスにより分類
St31
分類結果を、各分類結果のAの平均値により序列化
St32
隣合った分類結果w_1とw_2の境界値を
$B=\frac{m_2 \cdot d_1 + m_1 \cdot d_2}{d_1 + d_2}$
とする
m_i：Aの平均
d_i：Aの分数
St33
すべての分類結果をw_iに対して
w_i中のデータで、境界値にAの値が属さないものの割合があらかじめ決められたより大きいか？
St34
N
Y
w_iを２つに分割
St35
分割されたw_iがあったか？
St36
N
Y
分割された各区間に記号を付与
St37
記号化された特徴量出力
St38
END